Programación científica y análisis estadístico con Python

Master en Big Data

Sección 2: Pandas: Series y DataFrames

PROFESOR/A
Guillermo González Sánchez

Santander

www.eoi.es

Pandas

Pandas es una librería de Python con estructuras de datos de alto nivel y herramientas de procesamiento. Está diseñada para hacer el análisis de datos fluído.

Pandas está construída sobre Numpy.

Sus características fundamentales son:

- Estructuras de datos con etiquetado automático de ejes o especificando la alineación de los datos para prevenir errores.
- Funciones específicas para series temporales.
- Amplia funcionalidad de importación y exportación a distintos tipos de datos.
- Operaciones aritméticas y reducciones se pueden pasar directamente sobre los ejes, usando los índices.
- Manipulación flexible de datos incompletos.
- Operaciones de unión y otras relaciones esenciales de bases de datos relacionales.

PANel DAta S

Pandas

Pandas: Series y DataFrames

Los dos tipos de objetos con los que se trabaja en Pandas son

- Series:
  Es un objeto unidimensional, contiene un array de datos (de los tipos vistos en Numpy) y un array asociado de etiquetas para esos datos, llamado índice.

```
In [17]: s6 = pd.Series([65, 72, 65, 81, 56, 83, 61, 78, 65, 51],
                        index=['R0', 'R1', 'R2', 'R3', 'R4', 'R5', 'R6', 'R7', 'R8', 'R9'])
         s6

Out[17]: R0    65
         R1    72
         R2    65
         R3    81
         R4    56
         R5    83
         R6    61
         R7    78
         R8    65
         R9    51
         dtype: int64
```

- DataFrame:
  Representa los datos en una estructura similar a una hoja de cálculo. Contiene una colección ordenada de columnas pudiendo ser cada una de ellas de un tipo distinto de datos.

	Birth Month	Origin	Age	Gender
Carly	January	UK	27	f
Rachel	September	Spain	28	f
Nicky	September	Jamaica	28	f
Wendy	November	Italy	22	f
Judith	February	France	19	f

Un DataFrame contiene dos tipos de índice, para filas y para
columnas. Dentro de las posibilidades que hay de construir un
DataFrame, una de las más comunes es a partir de un diccionario Python con longitudes homogéneas de valores.

Series

La representación de un objeto Series muestra el índice en la parte izquierda y los valores en la derecha. Al final se muestra el tipo de datos del que está compuesto la serie.

```
In [3]: import pandas as pd

        obj = pd.Series([4, 7, -5, 3],['r1','r2','r3','r4'])
        obj

Out[3]: r1    4
        r2    7
        r3   -5
        r4    3
        dtype: int64
```

Pandas posee dos atributos para acceder tanto al índice como a los valores.

```
In [5]: obj.index

Out[5]: Index(['r1', 'r2', 'r3', 'r4'], dtype='object')

In [6]: obj.values

Out[6]: array([ 4,  7, -5,  3])
```

DataFrames

Los objetos DataFrame son tablas bidimensionales con índices de filas y columnas.

- Los índices son inmutables, esto hace las operaciones seguras.
- Los índices pueden tener múltiples niveles tanto en filas como en columnas. Se puede por tanto trabajar con un objeto de más de dos dimensiones en una representación de dos dimensiones.
- Se pueden seleccionar subconjuntos de la tabla con varias opciones de Pandas, así como aplicar funciones, agrupamientos, y variantes.

	year	state	pop	debt
one	2000	Ohio	1.5	NaN
two	2001	Ohio	1.7	NaN
three	2002	Ohio	3.6	NaN
four	2001	Nevada	2.4	NaN
five	2002	Nevada	2.9	NaN

	Estado	Ohio		Colorado
	Color	Green	Red	Green
Tipo	Número			
a	1	0	1	2
	2	3	4	5
b	1	6	7	8
	2	9	10	11

Agrupamiento y funciones

La estructura de agregación consiste en tres pasos secuenciales:
1. **División** - Dividir en grupos el DataFrame según valores en las columnas clave.
2. **Aplicación** - Aplicar una función a cada grupo resultante en la división.
Generalmente funciones agregantes, filtrantes ó transformantes.
3. **Combinación** - Combinar los resultados de los pasos anteriores en un nuevo DataFrame.